DOCUMENT RESUME

ED 247 258 TM 840 429

AUTHOR Oosterhof, Albert C.; Salisbury, David F.
TITLE Some Measurement and Instruction Related Considerations Regarding Computer Assisted Testing.
PUB DATE Apr 84
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Criterion Referenced Tests; Feedback; Higher Education; *Instructional Design; *Instructional Improvement; Remedial Instruction; *Test Construction; Test Reliability; Test Validity
IDENTIFIERS *Florida State University

ABSTRACT

The Assessment Resource Center (ARC) at Florida State University provides computer assisted testing (CAT) for approximately 4,000 students each term. Computer capabilities permit a small proctoring staff to administer tests simultaneously to large numbers of students. Programs provide immediate feedback for students and generate a variety of reports for instructors. Courses using the ARC for testing are required to provide an optional retest opportunity for students. Course objectives, test items, and related instructional materials have been developed by teams of instructors with an expert in testing and instructional development. The ARC experience has identified three measurement and two instructional design considerations crucial to the effectiveness of CAT operations in instructional environments: test quality; confidence in mastery decisions; maintenance of test validity; the quality of instructional objectives; and extended feedback following testing. The lack of effective methods to clarify test-identified learning deficiencies is the greatest weakness in the present ARC operation. Guidelines for producing and evaluating instructional objectives, test questions, and remediation procedures should be a central component of future CAT software documentation. Inappropriate computer applications to testing will have unfavorable effects on both the quality of instruction and the perceived benefits of educational measurement. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED247258

SOME MEASUREMENT AND INSTRUCTION RELATED CONSIDERATIONS REGARDING COMPUTER ASSISTED TESTING

Albert C. Oosterhof & David F. Salisbury

Florida State University

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

X Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. C. Oosterhof

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 429

SOME MEASUREMENT AND INSTRUCTION RELATED CONSIDERATIONS REGARDING COMPUTER ASSISTED TESTING

Albert C. Oosterhof & David F. Salisbury
Florida State University

As many have expected, advances in computer technology are having highly visible effects on instructional procedures in schools. Chambers and Bork (1980) reported that approximately 90% of a representative sample of 974 public school districts were using computers in instructional settings. It would probably be safe to say that most of these schools are utilizing the computer to perform at least some of the functions typically associated with computer assisted testing (CAT), including administering and scoring of tests, providing diagnostic feedback to students, presenting progress summaries to instructors, and providing statistical analysis of test items and other test analysis data.

Numerous descriptions of existing CAT systems have been presented (e.g. Dick & Gallager, 1972; Hammer & Henderson, 1972; Howze, 1978; McArthur & Choppin, 1982; Mitzel, 1974; Uffelman, 1978). The present discussion directs attention to several measurement and instruction considerations related to computer assisted testing. The concern of the authors is that many in the public schools and other instructional environments involved in the development and operation of CAT systems are not aware of these considerations. Consequently, this paper addresses several of these concerns within the context of an operational computer managed testing facility. The first part of the paper describes a CAT system developed for the Assessment Resource Center at Florida State University. The second part of the paper discusses measurement and instruction considerations we feel are particularly relevant to this system. Although the present CAT system is employed in a university setting, the basic attributes and concerns inherent in this system should generalize to other CAT environments.

Description of the Assessment Resource Center

The present system is used by the Assessment Resource Center (ARC) at Florida State University to provide computer assisted testing to approximately 4,000 students enrolled each term in courses offered by several academic departments. Each student comes to the ARC after completing a unit of instruction. The student fills in a machine scannable form indicating who he or she is and the identity of the unit of instruction on which the student is to be examined. The proctor inputs the form into a small optical scanner, which in turn transmits the information provided by the student to a computer. The computer accesses its files to determine if prerequisite units of instruction have been sufficiently mastered, and whether the student is maintaining an adequate rate of learning. If prerequisite conditions are satisfied, the computer identifies to the proctor the set of test materials to be provided the student.

TM 820 429

Upon receiving the test materials, the student takes an assigned seat within the center. Test materials are used by the student independent of a computer terminal, with student responses recorded on an answer form. After completing the test, the student returns the test materials and answer form to the proctor. The answer form is read by the optical scanner which relays the student's responses to the computer. The computer updates its files, and controls the printing of a prescriptive report which is handed to the student one or two minutes after the completion of the test. The student is expected to remedy difficulties diagnosed by the test and then return to the testing site within a predetermined amount of time for reassessment of these competencies. Students enrolled in a variety of courses are concurrently administered tests in the ARC. The management capabilities of a computer permit a small proctoring staff to administer tests simultaneously to a relatively large number of students. A series of report generation programs provide instructors with current summaries of student progress, estimates of final course grades, and analyses of test items.

Initiating Use of the ARC within a Course

The ARC was developed to facilitate improvements in both the quality of local test instruments and the use of test results. Because involvement of instructors was to be voluntary, the ARC was also to be responsive to needs typically perceived by an instructor or academic department. It was anticipated that instructors would be most interested in aspects of the system that increased time available for classroom instruction and facilitated management of student records.

The ARC developers sought the participation of what we refer to as "high-impact" courses. The impact of a course is a function of its enrollment, the degree to which the course serves as a feeder to subsequent courses, and the proportion of its students who are majors from other academic departments. Use of high enrollment courses permitted a larger number of students to benefit from a given amount of test development and other instructional design activities. It was anticipated that selecting courses which were prerequisite to other courses and which enrolled students from several academic departments would increase the visibility of the ARC and promote future involvement of other courses.

After a high-impact course was identified, the potential use of the ARC was described to its instructional staff in terms of benefits to be realized by the instructors and conditions required by the ARC developers. Benefits to the instructors emphasized the additional time for instruction that would result from moving testing external to the classroom, the reduced load on instructors provided by the student record management capability of the computer, the elimination of the need for arranging make-up exams due to the flexibility of the testing schedule in the ARC, and the increased monitoring and security possible through individually administered exams. The principal condition required by the ARC developers was that students have an option of taking a retest after receiving feedback from the initial test and reviewing course material suspected to be causing difficulty. This requirement was acceptable to instructors because ARC facilities allowed for retesting without additional involvement of the instructor or classroom, and there was a perceived advantage to students restudying material that was less effectively learned.

Once an instructor perceived merit to using the ARC and accepted the retesting condition, it was rather easy to communicate additional conditions that were logical requirements of the system. For example, feedback to students would not accurately identify areas of deficiency unless test items were carefully designed to measure these skills. Development of these test items would not be possible unless the specific skills to be measured were first identified in the form of course objectives. Providing feedback to students would not be meaningful unless a list of these skills was also provided to the members of the class, and meaningful references to study materials to be used between the initial test and the retest were provided students. These conditions are not new to either measurement specialists or instructional developers. However, simply providing a computer and relevant CAT software does not automatically result in application of these important conditions.

Course objectives, test items, and related instructional materials were produced by a team consisting of the instructors of the course and an additional member with testing and instructional development expertise. This team would establish course objectives, create test items which measured each of these objectives, and develop procedures to aid students in the review of content which was found to cause difficulty.

A uniform procedure was followed to complete each of the developmental tasks. The technical member would first describe qualities desired in the component being developed (e.g. objectives), the content members would develop initial examples, the technical member would review these examples, the technical and content members would jointly discuss the examples, the content members would revise the examples, and the technical member would review the revisions and make minor adjustments as appropriate. This interactive process would be repeated until objectives, test items, and study references were developed for a unit of instruction. Usually the content members would have other faculty from their department "validate" the objectives and test items after they were developed.

Some Measurement and Instruction Related Considerations

Three measurement and two instructional design issues were or have become dominant within the present application of computer assisted testing. It is our belief that individuals responsible for the development and dissemination of similar CAT systems should meaningfully respond to these issues if instructional gains are to be realized through the application of computer assisted testing procedures. The three measurement issues, in the order of their discussion, are: (a) quality of test instruments, (b) confidence in mastery decisions, and (c) maintenance of test validity.

Test Quality. The chapter written by Glaser and Nitko (1971) is but one of several presentations in the literature which addresses the important role testing plays within systematic instruction. Our experience with the ARC reinforces the opinion that classroom teachers are generally unfamiliar with even the most basic measurement considerations that are essential to the Glaser-Nitko model. For example, we observed that instructors are unaware of the importance of having test items sample a representative set of the skills being assessed. However, as stated

previously, we have found that the ARC or other computer assisted testing systems can facilitate instructors' realization of the need to follow good test construction principles. Once course objectives were identified and expressed clearly, it became obvious that computer generated student feedback reporting performance by objective would be meaningful only if test items are written to measure each of the specified skills. Under these conditions, instructors perceived and responded effectively to the need for content valid tests.

Computer assisted testing facilitates but does not inherently promote an improvement in test quality. To obtain interest from perspective users of the ARC, focus was initially placed on nonmeasurement features such as additional classroom time for instruction and automation of record keeping. This is where the focus appears to be with many CAT systems. But unless qualities critical to test construction are addressed, users of these systems may inappropriately anticipate improvement in learning. Use of the computer may actually make it more difficult for instructors to perceive the invalidity of tests. Computer printouts provide a sense of authenticity to test results---and if computer assisted testing removes testing from the classroom, as is the case with the ARC, it is more difficult for students to communicate irrelevancy when it exists in the tests.

Confidence in mastery decisions. As with the system used in the ARC, many computer assisted testing systems provide reports which list the performance of students on individual objectives. Typically, achievement on an objective is measured by a limited number of items, possibly as few as one. Millman (1974) used the binomial model to demonstrate that the probabilities of false-positive and negative decisions can be quite high unless substantial numbers of items are used to measure competence within each domain. For example, if a criterion of 70% proficiency is required, 25 items are required to reduce to 15% the number of students whose true score is 60% but who would obtain an observed score equal or greater than 18 (i.e., would obtain a score greater than 70% and would incorrectly be judged to have surpassed the criterion).

In most situations, locally developing and administering tests which contain as many as 25 items for each of several objectives is impractical. However, several alternatives are available. For instance, the reporting of results by individual objectives can be omitted and only aggregate scores provided. This eliminates the diagnostic value of the test (although one could argue the diagnostic value of the few items associated with each objective is minimal), and reduces substantially a potential value of computer assisted testing. A second alternative is to group objectives and associated objectives into broad domains, increasing the breath of a given domain until the minimum number of test items required for a specified degree of confidence is achieved. Although feedback by broad domains may be more useful to prescribing remediation than would be a single total score, it is often difficult to find meaningful groups of objectives.

A third alternative would be to question the amount of precision required in mastery decisions. One conceivably could reason that a distinction between 60% and 70% levels of performance is not of practical importance for many skills measured by written tests, but being able to

distinguish between the 40% and 70% proficiency levels is important. Again using the binomial model, only 4 items per objective are required to achieve a .18 probability of an individual with a true score of 40% correctly answering 3 or more of a sample of 4 items. And there is a .13 probability of an individual with a true score of 50% correctly answering 3 out of 3 items correctly, although the probability of obtaining false-negative classifications in this last example is increased significantly.

A fourth alternative is to administer tests that result in a reduced proportion of students scoring near the criterion level by increasing the variability of test scores. This can be accomplished by administering the test after an abbreviated period of instruction, by modifying the domain of items in order to increase their discrimination values, or by constructing items which maximally differentiate between degrees of achievement at the cutoff level of ability.

A fifth alternative is to simply accept high probabilities of misclassification if in fact it were found that specific diagnostic feedback with high error rates resulted on an average in more meaningful remedial activities (e.g. higher scores on a posttest) than if more general or no diagnostic information were provided. The criterion score might be raised to better control the number of false positives, if the increased number of false negatives could be tolerated. The appropriateness of the third, fourth, and fifth alternatives can be evaluated through research and discussions, and probably should not be rejected unconditionally.

Maintenance of test validity. A significant weakness of the ARC has been its inability to assure continued content and curricular validities of its tests. To date, this has been judged not to be a serious problem, but the potential clearly exists for it becoming an issue through inattention to existing test instruments used at the Center. We expect that this is a problem experienced by other CAT systems as well. Several factors contribute to the difficulty of maintaining content validity of tests. Existing forms of tests are updated as changes are made in textbooks, course content and course structure, but limits in personnel prevent detailed review of new tests. Changes in instructors can result in changes in emphasis within the curriculum without changes to course exams. To minimize the problem in the ARC, a review form is provided students along with test materials. The form asks students to comment about items they find unrelated to what is covered by the course, inconsistent with stated course objectives, or inappropriate for other reasons. Forms are forwarded to the instructor, but instructor use of these forms to improve tests is not monitored. Our problem with maintaining test validity is expected to increase with time.

Two instructional design issues that relate to computer assisted testing applications are to be discussed. These are: (a) quality of instructional objectives, and (b) the need for efficient and effective procedures for providing extended feedback following testing.

Quality of instructional objectives. The use of behavioral objectives has been both encouraged as a means of indicating to students what is expected of them and discouraged because of their potential for confining student learning to the subset of skills specified by the objectives. Melton (1978) references some empirical studies which suggest providing

students with behavior objectives enhances relevant learning, and other studies which fail to observe effects resulting from the use of the objectives. Melton proposes that degree of clarity is a major factor in determining whether or not specifying behavioral objectives increase learning of specified skills.

The clarity of objectives varies across courses administering tests in the ARC. In some courses, objectives could be characterized more as goal statements rather than as specification of desired behaviors. This causes difficulty in determining what is to be measured by test items, developing equivalent alternate forms to each test, and providing students with a clear statement of difficulties experienced in an exam. With increased experience, a greater proportion of the developmental effort was focused on preparing objectives. Instructors often felt less productive when writing objectives than when preparing test item. The instructors more readily accepted the importance of clearly written objectives when encouraged to prepare a very limited number of objectives before writing the initial test items.

Extended feedback following testing. In our judgment, the greatest weakness inherent with the present ARC operation is the lack of an efficient and effective means to clarify learning deficiencies identified by the tests. For security reasons, students are allowed to review tests only in the presence of instructors. Particularly for large enrollment courses, such a review is impractical. Although emphasis is placed on mastering objectives rather than the items sampling the behavior specified by an objective, a simple restatement of the objective is often insufficient for determining the cause of difficulty on a test. Giving a student the opportunity to review test items immediately following the administration of the test may clarify the specific nature of deficiencies to the student.

We propose that a CAT facility should include the capability of directing the learning of students subsequent to testing. This capability might include three components. The first component would allow the student to use onsite a test parallel to the one just completed. Items would be grouped by objective so that the student could focus on the types of items which caused the greatest difficulty. An appendix would provide the correct answer, and more importantly, the reason the answer was correct and the distractors incorrect. Recommendations by Gagne and Briggs (1979) provide a useful framework here. They describe feedback as one of the several essential events within instruction and indicate that the nature of the feedback should vary as a function of the type of learning involved. Ten types of learning are identified such as rule using, problem solving, and verbal information. For example in the case of rule using, the feedback to students should include an explanation of why the answer is correct or incorrect by referring back to the organizational rule. We believe this immediate feedback would in many cases help clarify the nature of the difficulty to the student.

The second component consists of one or more proctors being available to help students through problems with test items they are unable to solve by themselves. This is an approach similar to that taken by the personalized system of instruction (PSI) which has been widely adapted from ideas described by Keller (1968). To maintain efficiency, it is important

that this one-on-one assistance be used as prudently as possible. Proctors should be particularly familiar with the test items and explanations included in the first component so as to direct students back to their use when appropriate and as quickly as is practical. The proctors could have immediate access to students' answer sheets to tests just completed as well as worksheets when applicable. This could help students identify sources of problems undetectable from viewing the parallel items included in the first component. As with PSI, these proctors could be instructors, teaching assistants, or students who are interested in improving their own comprehension through providing explanations to others.

The third component would simply consist of an unsecured supply of test items which a student could later use as self tests. For example, after leaving the ARC, the student should review textual or other instructional material identified in the course syllabus as relevant to skills causing the greatest difficulty on the test. Whereas the immediate exposure to parallel items at the ARC may suggest specific causes of problems on a test, this review of instructional material is probably more likely to help the student broaden understanding of a concept beyond the highly specific behavior measured by a given test item. The availability of the unsecured test items could help reinforce central concepts, verify whether skills specified by objectives were comprehended, and build student confidence with these skills prior to taking a retest.

Summary

Implicit to each of the above considerations is the need to integrate measurement and instructional design principles into computer assisted testing systems. CAT systems are not instructionally meaningful unless test items relate to established objectives, the objectives communicate to students what is required of them, and an effective procedure is available to students for restudying material which caused difficulty on a test. This premise is not new, but appears to be in need of systematic application at the classroom level. Consequently, our principal recommendation is that individuals who are experienced in the application of measurement and instructional design principles associate themselves with developers and disseminators of CAT softwares and encourage the inclusion of guidelines for producing and evaluating instructional objectives, test questions, and remediation procedures as a central component within a system's documentation. As a form of professional service, measurement and instructional design specialists should continue to be encouraged to assist local users of computer assisted testing in the writing of clearly stated instructional objectives and content-valid test items. Where possible, resource centers should be institutionalized so that qualified staffs can assist teachers with the preparation of tests and other instructional materials that are to become an integral part of computerized testing procedures.

A computer assisted testing application such as the ARC has the potential of providing teachers with often needed clerical assistance and instructional time. Computer assisted testing also provides a vehicle through which the measurement and instruction related prerequisites can be easily communicated. Our experience suggests that if teachers are shown the logic of these prerequisites and are provided limited technical

support, they will become involved in meaningful instructional improvement activities independent of administrative directives. Cooley (1964), however, insightfully warned that computer-driven school testing programs would be of no educational significance unless educational technology was integrated into the advancing capabilities provided by computers. Twenty years later, this remains a most central concern. Unfortunately, inappropriate applications of computers to testing may not decrease usage of computer technology in education. But most certainly, such misuses of computers will have unfavorable effects on both the quality of instruction and on the perceived benefits of educational measurement.

References

Chambers, J. A., & Bork, A. (1980, September). Computer assisted learning in U.S. secondary and elementary schools. The Computing Teacher.

Cooley, W. W. (1964). A computer-measurement system for guidance. Harvard Educational Review, 34, 559-572.

Dick, W., & Gallager, P. (1972). Systems concepts and computer-managed instruction: an implementation and validation study. Educational Technology, 12, 33-39.

Gagne, R. M. & Briggs, L. J. (1979). Principles of instructional design (2nd ed.), New York: Holt, Rinehart, & Winston.

Glaser, R. & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Hammer, M., & Henderson, C. O. (1972). Improving large enrollment undergraduate instruction with computer generated repeatable testing. Proceedings of the 1972 Conference on Computers in Undergraduate Curricula, pp. 209-216. Atlanta, GA: Southern Regional Education Board (ERIC Document Reproduction Service No. ED 076 027).

Howze, G. (1978). An interactive software system for computer-assisted testing. AEDS Journal, 11, 31-37.

Keller, F. S. (1968). "Goodbye teacher...". Journal of Applied Behavior Analysis, 1, 79-89.

McArthur, D. L. & Choppin, B. (1982). Computerized individual testing and diagnosis. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Melton, R. F. (1978). Resolution of conflicting claims concerning the effect of behavioral objectives on student learning. Review of Educational Research, 48, 291-302.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.) Educational Measurement (2nd ed.). Berkeley: McCutchan.

Mitzel, H. E. (Ed.). (1974). Conference proceedings: An examination of the short range potential of computer-managed instruction (Contract No. NIE-C-74-0091). National Institute of Education.

Oosterhof, A. C. (1979). Cost-effective computer management of instruction appropriate for large-scale implementation. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Uffelman, R. L. (1978). Computer assisted testing at the education resource center. Paper presented at the annual meeting of the Association for the Development of Computer Based Instructional Systems, Dallas, TX.